Exploratory Data Analysis and Predictive Modeling for Diabetes Detection

Tommi Bimbato

May 7, 2025

1 Introduction

This document provides an overview of a project aimed at performing exploratory data analysis (EDA) and predictive modeling for diabetes detection. The project is part of the "Programming and Database - Data Science" course at the University of Verona (UNIVR), Academic Year 2024/2025.

It is important to note that this project is purely educational and should not be considered as having any medical or scientific validity. While it demonstrates the application of data science techniques in healthcare contexts, the results are intended solely for academic purposes and must not be used for professional diagnosis or treatment.

2 Project Overview

The dataset used, `diabetes_unclean.csv`, contains raw health data from various patients. The primary objectives are to clean this dataset, perform EDA, and develop predictive models to estimate diabetes likelihood in patients.

2.1 Data Cleaning and Exploration

The initial phase involves cleaning the dataset to handle missing values, outliers, and encoding categorical variables. This step ensures that the data is suitable for analysis and modeling.

- **Missing Values:** Rows with significant null values are removed.

- **Outliers:** Statistical thresholds based on physiological limits are applied to identify outliers, which are then handled appropriately.
- **Encoding:** Categorical variables such as gender and diabetes status are encoded into numerical values for modeling.

EDA is conducted using Jupyter notebooks. Key statistical summaries and visualizations (e.g., histograms, box plots) help identify patterns and correlations in the data.

3 Predictive Modeling

The next phase involves developing predictive models to estimate the likelihood of diabetes. The following models are used:

- Logistic Regression
- Decision Tree Classifier
- K-Nearest Neighbors (KNN)
- Random Forest Classifier

Each model is trained and tested using a train-test split approach, with special attention to handling class imbalance through random undersampling.

4 Integration and Web Application Development

The final phase involves integrating the models into a web application using Streamlit. This allows users to input patient data and receive real-time predictions on diabetes likelihood.

4.1 Streamlit Web App Features

- Readable interface showcasing data-cleaning, EDA, model selection and dataset balancing options.
- Real-time prediction from user input.
- Benchmarking feature to compare predictions across all models using the same user input data.

5 Structure of the Repository

- `data/`: Contains raw and processed datasets.
- `notebooks/`: Jupyter notebooks for data cleaning, EDA, and modeling.
- `src/`: Python scripts for model integration and web app development.

This document serves as a brief overview of the project's methodology and outcomes. For more detailed insights, refer to the Jupyter notebooks and code files in the repository.